

Some remarks on modularity density

Alberto Costa

Received: date / Accepted: date

Abstract A “quantitative function” for community detection called modularity density has been proposed by Li, Zhang, Wang, Zhang, and Chen in [Phys. Rev. E **77**, 036109 (2008)]. We study the modularity density maximization problem and we discuss some features of the optimal solution. More precisely, we show that in the optimal solution there can be communities having negative modularity density, and we propose a modification of the original formulation to overcome this issue. Moreover, we show that a clique can be divided into two or more parts when maximizing the modularity density. We also compare the solution found by maximizing the modularity density with that obtained by maximizing the modularity on the Zachary karate club network.

Keywords clustering · community detection · complex networks · modularity density maximization

1 Introduction

Networks, or graphs, are often used to describe complex systems, and they find application in many fields, e.g., biology and bioinformatics [14,18], recommender systems [1], social networks [12]. One of the topics related to networks which has been studied extensively in the last years is community detection: given a network $G = (V, E)$, where V is the set of vertices and E is the set of edges, one wants to find subsets of V (called clusters, or communities, or modules) which are more connected with vertices in the same community than with vertices in other communities. Hence, a partition is obtained by splitting

Financial support by SUTD-MIT International Design Center under grant IDG21300102.

A. Costa
Singapore University of Technology and Design
E-mail: costa@lix.polytechnique.fr

V in m communities $\{V_1, \dots, V_m\}$ which cover V . In general, these communities are non-empty, non-overlapping, and their number m is not known a priori.

There are many ways to define a community. For example, one may specify some rules that each community must respect [3, 4, 19]. Another approach is to use some heuristics (see for example [5, 12]). Alternatively, one could specify an objective function to maximize or minimize. Concerning the latter, probably the most famous of such functions is modularity, which represents the fraction of edges within communities minus the expected fraction of such edges in a random network with the same degree distribution [12, 17]. More precisely, using the notation of [15], modularity is defined as follows:

$$Q = \sum_{i=1}^m \left[\frac{L(V_i, V_i)}{L(V, V)} - \left(\frac{L(V_i, V)}{L(V, V)} \right)^2 \right], \quad (1)$$

where $L(V_i, V_i)$ is twice the number of edges in the community V_i , $L(V, V)$ is twice the number of edges of G (i.e., $2|E|$), and $L(V_i, V)$ is equal to the sum of degrees of vertices belonging to the community V_i . Notice that, in order to find a good quality partition, modularity should be maximized.

Although modularity is widely used, it presents some issues, as degeneracy and resolution limit [11, 13]. Degeneracy is related to the possible presence of several high modularity partitions which makes it difficult to find the global optimum. Resolution limit refers to the sensitivity of modularity to the total number of edges in the network, hence small communities may not be identified and remain hidden inside larger ones. To overcome the resolution limit of modularity, a measure called modularity density has been proposed by Li, Zhang, Wang, Zhang, and Chen in [15]. More precisely, modularity density is defined as:

$$D = \sum_{i=1}^m d(G_i) = \sum_{i=1}^m \left[\frac{L(V_i, V_i) - L(V_i, \bar{V}_i)}{|V_i|} \right], \quad (2)$$

where $d(G_i)$ is the modularity density associated with the community V_i , $L(V_i, \bar{V}_i)$ is the number of edges joining a vertex in V_i to a vertex belonging to another community, and $|V_i|$ is the number of vertices belonging to V_i .

This paper is organized as follows: in Section 2 we discuss some properties of the modularity density. In particular, in Section 2.1 we show that in the optimal solution there can be communities having a negative modularity density value, and we propose a constraint to overcome this issue. We also show how this constraint can help to derive a mixed integer linear programming reformulation of the problem, and we point out the relationship between this constraint and the weak definition of Radicchi *et. al* [19]. In Section 2.2 we show that a clique can be split in the optimal solution. After that, in Section 3 we comment some wrong and inaccurate statements of [15]. Finally, in Section 4 we present the conclusions.

2 Discussion on the properties of modularity density

We discuss in the following some features of modularity density.

2.1 Lower bound for modularity density of a community

As for modularity, one should maximize the modularity density to find a good quality partition. In fact, Li *et. al* [15] state that “clearly the maximum D value is often achieved when the network is correctly partitioned”. Intuitively, the modularity density of each community should assume a high value, but there are cases where some communities can have negative modularity density value in the optimal solution. To show this, consider the network with 31 vertices of Fig. 1: it consists of 7 cliques, each of them having 4 vertices (square shape), connected to a smaller clique with 3 vertices (circle shape). The optimal

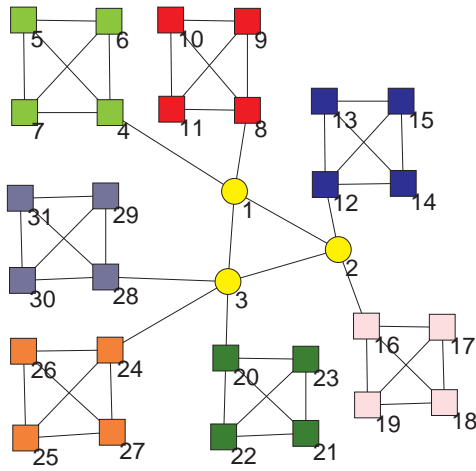


Fig. 1 Example of a network for which the optimal solution contains a community with negative modularity density (color online).

solution we found by solving the modularity density maximization problem using the exact formulations presented in [8] is a partition into 8 communities: 7 communities correspond to the 7 cliques having size 4 (i.e., the communities are $V_i = \{4i, 4i+1, 4i+2, 4i+3\}$, $\forall i \in \{1, \dots, 7\}$) whereas the other community corresponds to the smaller clique with 3 vertices (i.e., $V_8 = \{1, 2, 3\}$). It could be easily checked that the modularity density D of this partition is 18.9167. More precisely, the modularity density value associated with the community V_8 is $-\frac{1}{3}$, and $\frac{11}{4}$ for each of the other 7 communities. Hence, we cannot assume

that in the optimal solution each community has a non-negative modularity density value. Notice that this property is strictly related to the weak definition suggested by Radicchi *et al.* [19]. We discuss now this point more in detail. Let k_i^{in} be the number of edges connecting the vertex v_i to other vertices in the same community, and k_i^{out} be the number of edges connecting the vertex v_i to vertices belonging to other communities (hence, the degree of the vertex v_i is $k_i = k_i^{in} + k_i^{out}$). According to Radicchi *et al.* [19], a subgraph V_l of V is a community in the weak sense if:

$$\sum_{v_i \in V_l} k_i^{in} > \sum_{v_i \in V_l} k_i^{out}, \quad (3)$$

which implies that twice the number of edges inside a community is *strictly* greater than the number of edges connecting a vertex of the community to a vertex in another community (cut edges). Let x_{il} be a binary variable equal to 1 if the vertex v_i is inside the community l , and 0 otherwise, and let a_{ij} be an element of the adjacency matrix of the graph G (i.e., a_{ij} is equal to 1 if and only if there is an edge connecting v_i to v_j). As shown in [4], the weak condition (3) is equivalent to:

$$4 \sum_{\{v_i, v_j\} \in E} x_{il} x_{jl} \geq \sum_{v_i \in V} k_i x_{il} + 1. \quad (4)$$

According to Appendix A of [15], modularity density can be expressed as follows:

$$D = \sum_{l=1}^m \left(\frac{\sum_{v_i \in V} \sum_{v_j \in V} a_{ij} x_{il} x_{jl} - \sum_{v_i \in V} \sum_{v_j \in V} a_{ij} x_{il} (1 - x_{jl})}{\sum_{v_i \in V} x_{il}} \right), \quad (5)$$

which can be rewritten as:

$$D = \sum_{l=1}^m \left(\frac{4 \sum_{\{v_i, v_j\} \in E} x_{il} x_{jl} - \sum_{v_i \in V} k_i x_{il}}{\sum_{v_i \in V} x_{il}} \right). \quad (6)$$

Comparing (4) and (6) it appears that the weak definition is respected if, for each community, the corresponding value of modularity density is *strictly* positive. Therefore, one could adjoin to the modularity density formulation the constraint (4) without the +1 on the right-hand side to assure that each community has got a non-negative modularity density value, or the constraint (4) to assure that the partition found is compatible with the weak definition of [19]. The latter has been studied in [4] for modularity maximization. Let

$M = \{1, \dots, m\}$ be the set of the indices of the communities. The binary non-linear formulation which includes the weak constraint can be written as:

$$\max \sum_{l \in M} \left(\frac{4 \sum_{\{v_i, v_j\} \in E} x_{il} x_{jl} - \sum_{v_i \in V} k_i x_{il}}{\sum_{v_i \in V} x_{il}} \right) \quad (7)$$

$$\text{s.t. } \forall l \in M \quad 1 \leq \sum_{v_i \in V} x_{il} \leq |V| - 1 \quad (8)$$

$$\forall v_i \in V \quad \sum_{l \in M} x_{il} = 1 \quad (9)$$

$$\forall l \in M \quad 4 \sum_{\{v_i, v_j\} \in E} x_{il} x_{jl} \geq \sum_{v_i \in V} k_i x_{il} + L \quad (10)$$

$$\forall l \in M, \forall v_i \in V \quad x_{il} \in \{0, 1\}, \quad (11)$$

where (8) ensures that each community is non-empty and that all the vertices are not assigned to the same community (we suppose that there are at least two communities, otherwise the solution would be the trivial partition containing all the vertices), (9) imposes that each vertex belong to only one community, and (10) is the weak constraint, where the value of L is 1 if we consider the original definition in [19] and 0 if we only want to guarantee that each community assumes a non-negative value of modularity density.

The advantage of this new formulation, which will be discussed in the following, is that we can derive a more efficient exact linearization of the objective function. As noticed in [8], the difficult part is the linearization of the fractions arising in (7) (the products $x_{il} x_{jl}$ involving two binary variables can be easily linearized exactly using the Fortet inequalities [10] or the dual approach presented in [9]). To ease the explanation, we consider the modularity density of the community V_l (the same technique can be applied to linearize the modularity density of all the other communities). The idea used in [8] for the linearization the modularity density of V_l (formulation MDL) is the following. First, we introduce a variable α_l representing the modularity density of V_l :

$$\alpha_l = \frac{4 \sum_{\{v_i, v_j\} \in E} x_{il} x_{jl} - \sum_{v_i \in V} k_i x_{il}}{\sum_{v_i \in V} x_{il}}. \quad (12)$$

Thanks to the fact that empty communities are not allowed (see constraint (8)), the denominator of (12) is greater than 0, hence we can write:

$$4 \sum_{\{v_i, v_j\} \in E} x_{il} x_{jl} - \sum_{v_i \in V} k_i x_{il} = \sum_{v_i \in V} \alpha_l x_{il}. \quad (13)$$

We need now to linearize each product $\alpha_l x_{il}$. We can derive an exact linearization by means of the McCormick inequalities [16], because x_{il} is binary.

However, we need a lower and an upper bound on α_l . Indeed, the tighter those bounds, the better the linearization. Concerning the upper bound, it has been computed in [8] by solving an auxiliary problem, whereas a theoretical lower bound $L_\alpha = -\frac{k_{\max_1} + k_{\max_2}}{2}$ has been employed (where k_{\max_1} and k_{\max_2} are two vertices with the highest degrees). If constraint (10) holds then the lower bound for α_l would be $L = 1$ or $L = 0$ (depending on the value of L in (10)). Those values provide in general a lower bound which is much tighter than $L_\alpha = -\frac{k_{\max_1} + k_{\max_2}}{2}$, and which does not depend on the size of the instances (on the other hand, the quality of the bound $L_\alpha = -\frac{k_{\max_1} + k_{\max_2}}{2}$ decreases with the size of the instance, in general). This idea can be also extended to the formulation MDB in [8], where a binary decomposition of the denominator of (12) has been employed to decrease the number of products to linearize with the McCormick inequalities.

Using the formulation (7)-(11) with both $L = 0$ and $L = 1$, the partition into 8 communities represented in Fig. 1 is infeasible. The optimal solution is found when the number of communities is 7: the difference with respect to the previous case is that the clique $\{1, 2, 3\}$ is in the same community of one of the cliques having size 4 connected to vertex 3 (which one does not matter, the solution would be symmetric). The modularity density value associated with this new partition is 18.5.

2.2 Splitting of a clique in the optimal solution

Consider the network with 18 vertices presented in Fig. 2.

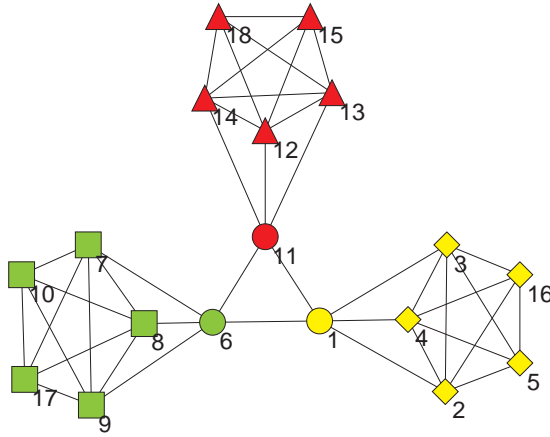


Fig. 2 Example of a network composed by three cliques with 5 vertices each, connected to a smaller clique having 3 vertices (color online).

Maximizing the modularity density when the number of communities is 4 results in a partition consisting of the 4 cliques: $V_1 = \{1, 6, 11\}$ (circle shape), $V_2 = \{2, 3, 4, 5, 16\}$ (diamond shape), $V_3 = \{7, 8, 9, 10, 17\}$ (square shape), $V_4 = \{12, 13, 14, 15, 18\}$ (triangle shape), and the corresponding modularity density value D is equal to 9.2. However, a higher value of modularity density is obtained when the number of communities is 3. The partition found is the following: $V_1 = \{1, 2, 3, 4, 5, 16\}$ (yellow color), $V_2 = \{6, 7, 8, 9, 10, 17\}$ (green color), $V_3 = \{11, 12, 13, 14, 15, 18\}$ (red color), and the corresponding value of modularity density is $D = 12$. Hence, in the optimal solution the small clique $\{1, 6, 11\}$ is split among the three other cliques. Notice that the solution with 4 communities would have been infeasible if using the formulation with the weak constraint (7)-(11), regardless of the value of L .

3 Comment on “Quantitative function for community detection”

Among the properties presented by Li *et al.* [15], some of them are not proved, wrong, or need to be clarified. Discussing and commenting these properties is the subject of this section.

3.1 Non-negative modularity density

Li *et al.* claim that “Since our purpose is to maximize the modularity density D , every term $d(G_i)$ must be non-negative”. Indeed, this is intuitive, as one may expect that the maximum value of D is obtained when all the terms $d(G_i)$ assume high values. Nevertheless, this is not always true when the number of communities is non-optimal (where the optimal number of communities is that of the partition yielding the highest value of modularity density). Consider for example the journal index network tested in Section V. 3 of [15]. The optimal number of communities is 4. However, when trying to maximize the modularity density with 5 communities, the authors state that “When we intend to split the network into five modules, we get essentially the same partition as with four, only with the singly connected journal Conservation Biology split off by itself as a community”. It is easy to check that the modularity density value of the community consisting only of the vertex associated to the journal Conservation Biology is -1. Actually, even when the number of communities is optimal, the property could not hold: in some cases having a community with a small negative value of modularity density allows other communities to assume higher modularity density values, thus yielding a higher value of D , as shown in Section 2.1. Notice that this wrong statement can yield wrong formulations for the modularity density maximization problem. As pointed out in Section 2.1, in [8] some exact linearizations of the non-linear formulation proposed in [15] are introduced, and they require a lower bound on the modularity density value. Using 0, as suggested in [15], would produce a wrong model. Moreover, the statement “the partition (subgraphs) by optimizing D results in communities

consistent with the weak definition suggested by Radicchi *et al.*” is also not correct. To summarize, there are two mistakes in their statements:

- it is not true that modularity density for a community is always non-negative in the optimal solution (as shown in Fig. 1);
- even though modularity density was non-negative for all communities in the optimal solution, this would not be enough to assure that the weak condition holds, because that condition requires the modularity density to be *strictly* positive for all communities (this because of the strict inequality in (3), that yields the +1 on the right-hand side of (4)).

3.2 Division of cliques in the optimal solution

One of the properties presented by Li *et al.* is “Given a clique with n vertices, maximizing modularity density or D does not divide it into two or more parts”. This statement should be clarified: the proof of the authors assumes the whole network being a clique (i.e., the clique has no external edges connecting it with other vertices), and it does not refer to any clique which can be found in a network (even though this property is later employed to prove some other results for networks containing some cliques, see Fig. 1 and Sections III. B-C in [15]). In fact, if the clique is densely connected to external vertices, it could be split, as shown in Section 2.2

3.3 Complexity of modularity density maximization

Li *et al.* state that “The search for optimal modularity density D is a **NP**-hard problem due to the fact that the space of possible partitions grows faster than any power of system size”. This is not an appropriate definition of **NP**-hardness. Consider for example the shortest path problem [7]: even though the space of the possible solutions is exponential, the problem belongs to **P**. This does not mean that modularity density maximization is not a **NP**-hard problem, but the correctness of this statement should be proven in a more appropriate way, for example by means of a reduction from a **NP**-complete problem to the decision version of the modularity density maximization (as done for modularity [2]). Notice that some papers already cite [15] as reference for the **NP**-hardness of modularity density maximization [6, 21].

3.4 Wrong result for Zachary karate club network

In Section V Li *et al.* test their function with some artificial and real-world instances. Concerning the latter, they present the results for the famous Zachary karate club network [20]. Commenting on the solution found, the authors claim that “By using our method, the network was partitioned into two communities

exactly consistent with real partition when $k = 2$ (see Fig. 3). However, maximizing the D value, we obtained the “optimal” partition with $k = 4$ which is also reasonable from the topology of the network”. We now discuss more in detail this point. We show in Fig. 3 (that is Fig. 3 borrowed from [15]) the partitions into 2 and 4 communities presented by the authors, and in Fig. 4 the same partitions with, in addition, the indications of the labels for the vertices.

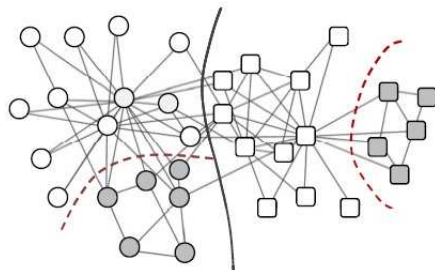


Fig. 3 Partitions into 2 and 4 communities of the Zachary karate club network presented in [15] (color online).

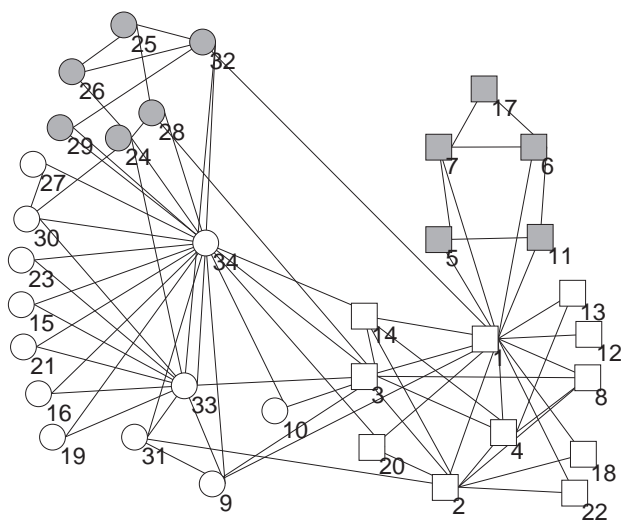


Fig. 4 Same partitions represented in Fig. 3 with labels for the vertices (color online).

We tried to optimize the D value on the Zachary karate club network, but we obtained different results. The partition with 2 communities found by Li

et al. consists of $V_1 = \{1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 17, 18, 20, 22\}$ (squares) and $V_2 = \{9, 10, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34\}$ (circles). When the number of communities is 4, each community of the previous partition is further split in two. The result is a partition composed by $V_1 = \{5, 6, 7, 11, 17\}$ (dark squares), $V_2 = \{1, 2, 3, 4, 8, 12, 13, 14, 18, 20, 22\}$ (white squares), $V_3 = \{24, 25, 26, 28, 29, 32\}$ (dark circles), and $V_4 = \{9, 10, 15, 16, 19, 21, 23, 27, 30, 31, 33, 34\}$ (white circles). We solved the problem of modularity density maximization for the Zachary karate club network with 2, 3, and 4 communities using the exact formulation presented in [8]. The result obtained with 2 communities is consistent with that of the authors, and the value of D is 6.83333. The result obtained with 3 communities is the following partition: $V_1 = \{1, 2, 3, 4, 8, 10, 12, 13, 14, 18, 20, 22\}$, $V_2 = \{9, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34\}$, and $V_3 = \{5, 6, 7, 11, 17\}$, with $D = 7.8451$. Notice that this is the same partition obtained with the almost-strong rule [3]. Finally, the partition into 4 communities gave a different result from that of the authors. The solution we found is the following: $V_1 = \{5, 6, 7, 11, 17\}$, $V_2 = \{1, 2, 3, 4, 8, 12, 13, 14, 18, 20, 22\}$, $V_3 = \{25, 26, 29, 32\}$, and $V_4 = \{9, 10, 15, 16, 19, 21, 23, 24, 27, 28, 30, 31, 33, 34\}$, with a value of D of 7.54481. The difference with respect to the solution of the authors is that vertices 24 and 28 are moved from the community V_3 to V_4 . Actually, if vertices 24 and 28 belong to V_3 , the corresponding value of D is 7.50909, that is non-optimal when there are 4 communities. These results are summarized in Table 1, together with the values of modularity Q associated with the partition found by maximizing the modularity density. Looking at Table 1, we can also notice that the solution

Table 1 Results obtained by maximizing the modularity density D on the Zachary karate club network, and corresponding values of modularity Q . The results refer to the optimal partitions obtained with 2, 3, and 4 communities, as well as to the non-optimal partition into 4 communities presented by the authors in [15] (see Fig. 3-4).

m	D	Q
2	6.83333	0.371466
3	7.8451	0.402038
4	7.54481	0.415105
4 (Fig. 3-4)	7.50909	0.41979

with 4 communities of Fig. 3 corresponds to the highest value of modularity Q , whereas the best solution found by maximizing D with 4 communities is different, as explained earlier. To summarize, not only the solution proposed in [15] is non-optimal with respect to the number of communities (which should be 3 instead of 4), but even fixing the number of communities at 4 their solutions is non-optimal with respect to the modularity density value (which should be 0.754481 instead of 0.750909). The reason for this behavior is that all the results presented in [15] are based on a method which finds only local optima (as confirmed by the authors), hence there is no guarantee of global optimality.

4 Conclusion

In this paper we have discussed some properties of modularity density, and we have shown the relationship with the weak definition of community of Radicchi *et. al* [19]. This remark allowed us to derive a new formulation, which is easier to linearize and which ensures that in the optimal solution each community has a non-negative value of modularity density. Moreover, we have clarified, commented, and corrected some wrong and inaccurate statements presented in [15]. Despite these issues, modularity density remains a very interesting criterion, due to its capability of fixing the resolution limit issue of modularity. For this reason, we targeted our effort to a better characterization and description of its features.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* **17**(6), 734–749 (2005)
2. Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering* **20**(2), 172–188 (2008)
3. Cafieri, S., Caporossi, G., Hansen, P., Perron, S., Costa, A.: Finding communities in networks in the strong and almost-strong sense. *Physical Review E* **85**(4), 046,113 (2012)
4. Cafieri, S., Costa, A., Hansen, P.: Adding cohesion constraints to models for modularity maximization in networks. *Journal of Complex Networks* (2014). Accepted
5. Cafieri, S., Costa, A., Hansen, P.: Reformulation of a model for hierarchical divisive graph modularity maximization. *Annals of Operations Research* **222**(1), 213–226 (2014)
6. Chen, M., Kuzmin, K., Szymanski, B.: Community detection via maximization of modularity and its variants. *IEEE Transactions on Computational Social Systems* **1**(1), 46–65 (2014)
7. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 3rd edn. The MIT Press (2009)
8. Costa, A.: MILP formulations for the modularity density maximization problem. Tech. Rep. 2014-10-4588, Optimization Online (2014)
9. Costa, A., Liberti, L.: Relaxations of multilinear convex envelopes: Dual is better than primal. In: R. Klasing (ed.) *Experimental Algorithms, Lecture Notes in Computer Science*, vol. 7276, pp. 87–98. Springer Berlin Heidelberg (2012)
10. Fortet, R.: Applications de l’algèbre de Boole en recherche opérationnelle. *Revue Française de Recherche Opérationnelle* **4**(14), 17–26 (1960)
11. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the U.S.A.* **104**(1), 36–41 (2007)
12. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the U.S.A.* **99**(12), 7821–7826 (2002)
13. Good, B.H., de Montjoye, Y.A., Clauset, A.: Performance of modularity maximization in practical contexts. *Physical Review E* **81**(4), 046,106 (2010)
14. Guimerà, R., Amaral, L.A.N.: Functional cartography of complex metabolic networks. *Nature* **433**, 895–900 (2005)
15. Li, Z., Zhang, S., Wang, R.S., Zhang, X.S., Chen, L.: Quantitative function for community detection. *Physical Review E* **77**, 036,109 (2008)
16. McCormick, G.: Computability of global solutions to factorable nonconvex programs: Part I — Convex underestimating problems. *Mathematical Programming* **10**, 146–175 (1976)

17. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. *Physical Review E* **69**(2), 026,113 (2004)
18. Palla, G., Dernyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**(7043), 814–818 (2005)
19. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the U.S.A.* **101**(9), 2658–2663 (2004)
20. Zachary, W.: An information flow model for conflict and fission in small group. *Journal of Anthropological Research* **33**, 452–473 (1977)
21. Zhang, S., Ning, X., Ding, C.: Maximizing Modularity Density for Exploring Modular Organization of Protein Interaction Networks. In: *The Third International Symposium on Optimization and Systems Biology (OSB'09)*, *Lecture Notes in Operations Research* 11, pp. 361–370 (2009)